Aplicación de la Minería de Datos para el análisis de información clínica. Estudio Experimental en cardiopatías isquémicas

Application of Data Mining for analysis of clinical information. Experimental study on coronary heart disease

M.Sc. Ingrid Wilford Rivera. Facultad de Ingeniería Informática, Instituto Superior Politécnico José Antonio Echeverría, Cuba. Dirección postal: Ave. 59 #11416 e/ 114 y 116, Marianao, CP 11400, Ciudad de La Habana, Cuba. Email: iwilford@ceis.cujae.edu.cu

Dr. Alejandro Rosete Suárez. Facultad de Ingeniería Informática, Instituto Superior Politécnico José Antonio Echeverría, Cuba

Lic. Alfredo Rodríguez Díaz. Centro para el Desarrollo Informático en la Salud, MINSAP, Cuba

Resumen

En Cuba, como resultado de la atención a pacientes que padecen cardiopatía isquémica se generan grandes volúmenes de información anualmente. El análisis exhaustivo de dicha información tiene un gran valor de cara a las investigaciones científicas en cardiología. En este trabajo se expone un ejemplo de aplicación de la minería de datos para el apoyo a la toma de decisiones en esta especialidad de la medicina, a partir del estudio de las coronariografías realizadas a pacientes con cardiopatía isquémica.

Palabras Claves: Minería de Datos, Cardiología, Coronariografía, Cardiopatía Isquémica.

Abstract:

In Cuba, large amounts of information are generated every year as a result of the treatment of patients with ischemic heart disease. A thorough analysis of this information has great value for the scientific research in cardiology. This paper presents an example of applying data mining techniques to support the decision making process in this specialty. It exposes the results of the study of the data from coronary angiographies performed in patients with ischemic heart disease.

Key Words: Data mining, Cardiology, Coronary Angiography, Ischemic Heart Disease

Introducción

Existen numerosos dominios donde la Minería de Datos puede ser aplicada, en principio en todas las áreas o actividades que generen datos. Desde la década de los años 90, del pasado siglo, se vienen aplicado intensamente técnicas de minería de datos con diversos fines: apoyo a la toma de decisiones, gestión de procesos industriales, investigación científica, soporte al diseño de bases de datos y mejora de la calidad de los datos, entre otros.

El análisis de los datos de una base de datos, en ocasiones, se realiza mediante consultas efectuadas con lenguajes como el SQL (*Structured Query Language*), por lo que se produce sobre una base de datos operacional, es decir, junto al procesamiento transaccional en línea (*On-Line Transaction Processing*, OLTP), de las aplicaciones. Esta forma de análisis de datos sólo permite generar información resumida de una manera previamente establecida, poco flexible y poco escalable a grandes volúmenes de datos [1].

La tecnología de bases de datos ofreció posteriormente una arquitectura: el almacén de datos (*data warehouse*), que consiste en un repositorio de fuentes de datos heterogéneos, integrados y organizados bajo un esquema unificado para facilitar su análisis y dar soporte a la toma de decisiones. La tecnología de *data warehouse* soporta operaciones de procesamiento analítico en línea (*On-Line Analytical Processing*, OLAP), es decir, técnicas de análisis descriptivo y de sumarización, como pueden ser el resumen, la consolidación o la agregación, además de la posibilidad de ver la información desde distintas perspectivas. Sin embargo, esta tecnología no permite obtener reglas, patrones, pautas, es decir, conocimiento que pueda ser aplicado a nuevos datos [1].

Existe en la actualidad un conjunto de herramientas y técnicas que soportan la extracción de conocimiento útil a partir de los datos disponibles, y que se agrupan bajo el calificativo de "minería de datos".

Minería de Datos

En esencia la minería de datos (*data mining*), es un mecanismo de explotación y análisis, que consiste en la búsqueda y extracción de información valiosa, patrones y reglas ocultos en grandes volúmenes de datos [2][3].

La minería de datos se distingue de las aproximaciones comentadas anteriormente, ya que como resultado no obtiene datos sino conocimiento. El resultado de la minería de datos son conjuntos de reglas, ecuaciones, árboles de decisión, redes neuronales, grafos probabilísticos, los que pueden usarse para, por ejemplo, responder a interrogantes como: ¿existe un grupo de pacientes que se comporta de manera diferenciada? o ¿existen asociaciones entre los factores de riesgos presentes en los pacientes que padecen cierta enfermedad?.

La minería de datos resulta muy útil en situaciones donde el volumen de datos es muy grande o complejo por la cantidad de variables que se manipulan, o donde los especialistas no están disponibles para el análisis de los datos y la extracción de conocimiento. La tecnología de minería de datos posee valiosas cualidades:

- Descubrimiento automatizado de modelos previamente desconocidos
- Predicción automatizada de tendencias
- Procesamiento más rápido

La minería de datos contribuye a la toma de decisiones tácticas y estratégicas, proporcionando un sentido automatizado para identificar información clave, desde grandes volúmenes de datos generados por procesos tradicionales.

Minería de Datos Distribuida

En muchísimos dominios de aplicación, los datos se encuentran distribuidos en varios nodos ubicados en sitios distantes. Los avances en la informática y las comunicaciones han favorecido el desarrollo de este tipo de aplicaciones. En estos casos, por lo general, no es posible o factible centralizar todos los datos del sistema de información distribuido en un único repositorio, con el propósito de realizar tareas de minería de datos, debido, por ejemplo, a restricciones económicas, técnicas o legales. Por ello es necesario aplicar técnicas de minería de datos sobre múltiples fuentes o almacenes de datos. La minería de datos sobre fuentes de datos distribuidas se denomina minería de datos distribuida (*distributed data mining*).

Las bases de datos distribuidas (BDD) se pueden clasificar en homogéneas o heterogéneas. Las BDD homogéneas son aquellas en las que el mismo esquema está repetido en cada servidor y son, por lo tanto, los registros los que se encuentran repartidos en los diferentes nodos. Mientras que, las BDD heterogéneas son aquellas en las que cada parte o nodo almacena un subconjunto de las tablas o incluso atributos diferentes de una misma tabla.

Como resultado de la aplicación de las técnicas de minería de datos en fuente de datos homogéneas distribuidas se obtiene, para cada fuente de datos, un modelo o conjunto de patrones válidos localmente, es decir para la fuente de datos correspondiente, pero no, necesariamente, para el conjunto formado por todos los datos distribuidos (como si estos estuviesen centralizados en un mismo repositorio). Por lo tanto, será preciso integrar los diferentes modelos locales para obtener un modelo global que represente patrones coherentes y válidos para el conjunto de todos los datos distribuidos. La integración de estos modelos locales no es una tarea sencilla, ya que los mismos pueden carecer de detalles necesarios para obtener un modelo global de calidad.

Estudio Experimental

A continuación se describe el estudio experimental realizado cuyo objetivo general es analizar, mediante técnicas de minería de datos, información clínica correspondiente a un conjunto de coronariografías realizadas a pacientes con cardiopatía isquémica. De este objetivo general se derivaron los siguientes objetivos específicos:

1. Determinar un modelo de clasificación de las coronariografías, tomando “Complicaciones” como variable objetivo (esta variable admite dos posibles valores o clases: ‘Sí’ o ‘No’).

2. Determinar las reglas de asociación más relevantes que existen entre todas las variables analizadas.
3. Cumplir los objetivos anteriores:
   a. Tomando como conjunto de entrenamiento todas las coronariografías en un único dataset (D). (Análisis de datos centralizados)
   b. Manipulando el conjunto de entrenamiento D para obtener dos datasets o subconjuntos de coronariografías disjuntos (D1 y D2), donde D1 U D2 = D, y aplicar las mismas técnicas de minería de datos a cada subconjunto por separado. (Análisis de fuentes homogéneas de datos disjuntos)
4. Comparar los modelos obtenidos en ambas variantes (3a. y 3b.).

Métodos

El estudio experimental realizado abarcó las siguientes fases: preparación de los datos, resolución o minería de datos y análisis de los resultados. Para la aplicación de las técnicas de minería de datos se utilizó la versión 3.5.5 de la herramienta Weka [4].

Preparación de los datos

En el conjunto de datos que se recuperó inicialmente se tenían para cada coronariografía realizada, un total de 54 variables. Sin embargo, después de estudiar las características de los datos registrados y de consultar a los especialistas, se decidió considerar en este estudio experimental, solamente doce de las variables registradas. Estas variables fueron: fecha de realización de la coronariografía, edad del paciente, sexo, hospital o provincia que remite al paciente, diagnóstico clínico preoperatorio, cantidad de infartos anteriores, factores de riesgo (hipercolesterolémico, hipertenso, fumador, diabético), técnica empleada en la coronariografía y tipo de complicación.

Una vez seleccionadas las variables se prosiguió con la realización de una serie de transformaciones necesarias para la obtención de los datos derivados requeridos en el análisis, conformando de esta forma los conjuntos de entrenamiento a utilizar (D, D1 y D2). El conjunto de entrenamiento D quedó conformado por un total de 2445 registros o instancias de coronariografías. A partir de este conjunto de entrenamiento se generaron los conjuntos disjuntos y homogéneos entre sí y respecto a D: D1 y D2, de 1222 y 1223 registros respectivamente. Para ello, se seleccionaron aleatoriamente los 1222 registros de D1, conformando, entonces, los 1223 registros restantes el conjunto de entrenamiento D2.

Análisis descriptivo

Con el propósito de estudiar la composición de los datos almacenados, previo a la aplicación de los algoritmos de minería de datos, se realizó un análisis descriptivo de cada conjunto de entrenamiento. Para ello, se utilizaron los recursos de visualización de la herramienta Weka. Como resultado se llegaron a conclusiones muy interesantes. A modo de ilustración se ejemplifican algunas de ellas correspondientes a cada conjunto de entrenamiento.

Conjunto de entrenamiento D:

- La variable “diagnóstico clínico preoperatorio” admite 52 posibles valores, de estos, apenas 2 son los que predominan en el conjunto de entrenamiento (Tabla 1).

Tabla 1 –Distribución de la variable "diagnóstico clínico preoperatorio" (Conjunto de datos D)

Diagnóstico clínico preoperatorio	Total	Porcentaje
Angina de esfuerzo	1288	52.4%
Angina de empeoramiento progresiva	494	20.2%
Otros	663	27.4%
Total	2445	100%

.

- De los pacientes a los que se le realizaron coronariografías en este período (1998-2006):

- El 76% es del sexo masculino, siendo similar, en ambos sexos, la proporción de coronariografías que presentaron complicaciones respecto al total.
- El 59% tiene entre 46 y 60 años de edad. Se observa una marcada diferencia entre estos pacientes y los que tienen entre 61 y 70 años de edad, en cuanto a la proporción de las coronariografías que presentaron complicaciones respecto al total, siendo mayor en los pacientes cuya edad está en el rango de los 61 a los 70 años.

Conjunto de entrenamiento D1:

- La variable "diagnóstico clínico preoperatorio" tiene una distribución similar que en el conjunto D (Tabla 2).

Tabla 2 –Distribución de la variable "diagnóstico clínico preoperatorio" (Conjunto de datos D1)

Diagnóstico clínico preoperatorio	Total	Porcentaje
Angina de esfuerzo	627	51.3%
Angina de empeoramiento progresiva	239	19.6%
Otros	356	29.1%
Total	1222	100%

- De los pacientes a los que se le realizaron coronariografías en este período (1998-2006):

- El 75% es del sexo masculino, siendo similar, en ambos sexos, la proporción de coronariografías que presentaron complicaciones respecto al total.
- El 63% tiene entre 46 y 60 años de edad. De la misma forma, se observa una marcada diferencia entre estos pacientes y los que tienen entre 61 y 70 años de edad, en cuanto a la proporción de las coronariografías que presentaron complicaciones respecto al total.

Tabla 3 –Distribución de la variable "diagnóstico clínico preoperatorio" (Conjunto de datos D2)

Diagnóstico clínico preoperatorio	Total	Porcentaje
Angina de esfuerzo	661	54.0%
Angina de empeoramiento progresiva	255	20.9%
Otros	307	25.1%
Total	1223	100%

Conjunto de entrenamiento D2:

- En la Tabla 3 se puede observar la distribución de la variable "Diagnóstico clínico preoperatorio" para este conjunto de entrenamiento.
- De los pacientes a los que se le realizaron coronariografías en este período (1998-2006):
  - El 76% es del sexo masculino, siendo similar, en ambos sexos, la proporción de coronariografías que presentaron complicaciones respecto al total.
  - El 55% tiene entre 46 y 60 años de edad. En este caso, en todos los rangos de edades es similar la proporción de coronariografías que presentaron complicaciones respecto al total.

Después de estudiar las características de cada conjunto de entrenamiento se prosiguió a ejecutar los algoritmos de minería de datos.

Algoritmos aplicados

Las técnicas de minería de datos que se aplicaron en el estudio experimental, de acuerdo a los objetivos específicos definidos, fueron: clasificación (algoritmo J48, implementado en Weka) y asociación (algoritmo Apriori, implementado en Weka). A continuación se describen ambos algoritmos.

El algoritmo J48 implementado en Weka [5] es una versión del clásico algoritmo de árboles de decisión C4.5 propuesto por Quilan [6]. Los árboles de decisión entran dentro de los métodos de clasificación supervisada, es decir, se tiene una variable dependiente o clase, y el objetivo del clasificador es determinar el valor de dicha clase para casos nuevos.

El proceso de construcción del árbol comienza por el nodo raíz, el que tiene asociados todos los ejemplos o casos de entrenamiento. Lo primero es seleccionar la variable o atributo a partir de la cual se va a dividir la muestra de entrenamiento original (nodo raíz), buscando que en los subconjuntos generados haya una mínima variabilidad respecto a la clase. Este proceso es recursivo, es decir, una vez que se haya determinado la variable con la que se obtiene la mayor homogeneidad respecto a la clase en los nodos hijos, se vuelve a realizar el análisis para cada uno de los nodos hijos. Aunque en el límite este proceso se detendría cuando todos los nodos hojas contuvieran casos de una misma clase, no siempre se desea

llegar a este extremo, para lo cual se implementan métodos de pre-poda y post-poda de los árboles.

El algoritmo J48 amplía las funcionalidades del C4.5, tales como permitir la realización del proceso de post-poda del árbol mediante un método basado en la reducción del error (*reducedErrorPruning*) o que las divisiones sobre las variables discretas sean siempre binarias (*binarySplits*) [4][5]. Algunas propiedades concretas de la implementación son las siguientes:

- Admite atributos simbólicos y numéricos, aunque la clase debe ser simbólica.
- Se permiten ejemplos con valores desconocidos.
- El criterio de división está basado en la entropía y la ganancia de información.

Por su parte, Apriori es un algoritmo de aprendizaje de reglas de asociación muy simple y popular que permite identificar las posibles correlaciones o interdependencias entre distintas acciones o sucesos, lo que posibilita reconocer cómo la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros [7]. Las reglas de asociación son una manera de expresar patrones de datos de una base de datos. Estos patrones pueden servir para conocer el comportamiento general del problema que genera la base de datos, y de esta manera, disponer de información que pueda asistir en la toma de decisiones. Una regla de asociación es una proposición probabilística sobre la ocurrencia de ciertos estados en una base de datos. A diferencia de las reglas de clasificación, en la parte derecha de las reglas de asociación puede aparecer cualquier atributo, y además puede aparecer más de un atributo.

Para evaluar la calidad de una regla de asociación, se suelen emplear dos medidas: soporte (o cobertura) y confianza (o precisión). El soporte de una regla se define como el número de instancias que la regla predice correctamente (1). Por otra parte, la confianza mide el porcentaje de veces que la regla se cumple cuando se puede aplicar, es decir, cuando se cumple su antecedente (2).

$$\text{soporte}(A \rightarrow B) = P(A \cap B) \quad (1)$$

$$\text{confianza}(A \rightarrow B) = P(B \mid A) = \frac{P(A \cap B)}{P(A)} \quad (2)$$

El funcionamiento del algoritmo Apriori [7] se basa en la búsqueda de los conjuntos de *items* que cumplen con determinado umbral de soporte. Para ello, en primer lugar se construyen los conjuntos formados por sólo un *item*, que superan el soporte mínimo. Posteriormente, estos conjuntos se utilizan para construir los conjuntos de dos *items*, y así sucesivamente hasta que se llegue a un tamaño en el que no existan conjuntos de *ítems* con el soporte requerido. Una vez que se han seleccionado los conjuntos de *items* que cumplen con el soporte mínimo, el siguiente paso consiste en generar, a partir de estos conjuntos, las reglas de asociación que tengan un nivel de confianza mínimo. Del conjunto de reglas generadas, las que resultan más interesantes son aquellas que tienen su valor de soporte más alto.

La versión del algoritmo Apriori que implementa la herramienta WEKA es ligeramente distinta al explicado anteriormente. Así, el algoritmo no obtiene de una vez todos los conjuntos de *items* frecuentes que cumplen con el valor de umbral, sino que va iterando y

cada vez se obtienen los de un tamaño determinado, y con estos va generando reglas. Además, para mejorar la eficiencia del algoritmo en la búsqueda de los conjuntos de *items* frecuentes, elimina los atributos que tengan valores desconocidos en todos los ejemplos. Por otra parte, el algoritmo Apriori de Weka permite seleccionar las reglas atendiendo a diferentes métricas no únicamente la confianza [6][7].

Resultados

A continuación se describen, para cada conjunto de entrenamiento, los resultados fundamentales de la aplicación de los algoritmos de minería de datos seleccionados.
Conjunto de entrenamiento D:
Los resultados más precisos y de mayor calidad se obtuvieron con un factor de confianza (parámetro del algoritmo J48 usado para la poda del árbol) de 50%, lográndose una clasificación correcta en el 88.2% de las instancias.
A partir del árbol de decisión se identificaron reglas de clasificación muy interesantes, como por ejemplo:

- Se realizaron 398 coronariografías (16.3%) a pacientes no hipertensos que habían tenido al menos un infarto previamente, y no presentaron ninguna complicación.
- Se realizaron 95 coronariografías (3.9%) a pacientes hipertensos, cuya edad estaba en el rango de los 61 a los 70 años, que habían tenido al menos un infarto previamente y cuyo diagnóstico preoperatorio fue “Angina de esfuerzo”, y sí presentaron algún tipo de complicación en la operación.

Por otra parte, para la ejecución del algoritmo Apriori se eliminó la variable objetivo del conjunto de entrenamiento y se indicó como factor de confianza mínimo el 90%. Como resultado de su ejecución se obtuvieron reglas de asociación que permitieron llegar a conclusiones muy interesantes, entre estas:

- En el 94% de las coronariografías realizadas a pacientes no hipertensos, estos tampoco eran diabéticos, lo que se verifica en el 35% de las coronariografías analizadas (864 de 2445).
- En el 91% de las coronariografías realizadas a pacientes no hipertensos, estos tampoco eran hipercolesterolémicos, lo que se verifica en el 34% de las coronariografías analizadas (832 de 2445).

Las reglas de clasificación y asociación identificadas no sólo permiten describir la muestra estudiada sino que posibilitan predecir el comportamiento de nuevos casos.
Conjunto de entrenamiento D1:
Los resultados más precisos y de mayor calidad se obtuvieron con un factor de confianza de 30%, lográndose una clasificación correcta en el 86.6% de las instancias.
A partir del árbol de decisión se identificaron reglas de clasificación muy interesantes, como por ejemplo:

- Se realizaron 603 coronariografías (49.3%) a pacientes que no habían tenido ningún infarto previamente, y no presentaron ninguna complicación.

En este caso, para la ejecución del algoritmo Apriori se eliminó, también, la variable objetivo del conjunto de entrenamiento y se indicó como factor de confianza mínimo el 85%. Como

resultado de su ejecución se obtuvieron reglas de asociación que permitieron llegar a conclusiones muy interesantes, entre estas:

- El 100% de las coronariografías realizadas a pacientes cuya edad estaba en el rango de los 46 a los 60 años y que no habían tenido ningún infarto previamente, no presentó ninguna complicación. Verificándose esto en el 35% de las coronariografías analizadas (429 de 1222).
- El 100% de las coronariografías en las que se empleó la técnica “Coronariografía por vía A.B.D.” realizada a pacientes no hipercolesterolémicos y no hipertensos, no presentó ninguna complicación. Verificándose esto en el 25% de las coronariografías analizadas (301 de 1222).

Conjunto de entrenamiento D2:
Los resultados más precisos y de mayor calidad se obtuvieron con un factor de confianza de 100%, lográndose una clasificación correcta en el 91% de las instancias.
A partir del árbol de decisión se identificaron reglas de clasificación muy interesantes, como por ejemplo:

- Se realizaron 388 coronariografías (31.7%) a pacientes, que no habían tenido ningún infarto previamente, cuya edad estaba en el rango de los 46 a los 60 años, y no presentaron complicaciones en la operación.

De la misma forma, para la ejecución del algoritmo Apriori se eliminó la variable objetivo del conjunto de entrenamiento y se indicó como factor de confianza mínimo el 85%. Como resultado de su ejecución se obtuvieron reglas de asociación que permitieron llegar a conclusiones muy interesantes, entre estas:

- El 100% de las coronariografías realizadas a pacientes no hipertensos y que no habían tenido ningún infarto previamente, no presentó ninguna complicación. Verificándose esto en el 22% de las coronariografías analizadas (264 de 1223).

Análisis comparativo

En el estudio experimental realizado se observó que el tiempo que demoran los algoritmos (J48 y Apriori) en construir el modelo de minería de datos es inferior cuando la cantidad de registros del conjunto de entrenamiento es menor. La razón en la que disminuye dicho tiempo, en general, es similar a la razón en la que disminuye la cantidad de registros del conjunto de entrenamiento. Este hecho resulta muy interesante y puede tener implicaciones importantes en el desarrollo de nuevas versiones de algoritmos de minería de datos paralelos.
Por otra parte, se analizó en cada caso la semejanza, en cuanto al conocimiento descubierto, entre los modelos obtenidos a partir de los diferentes conjuntos de entrenamiento. Esto se hizo con el propósito de valorar la posibilidad de obtener, a través de la “integración” de los modelos resultantes del análisis de D1 y D2 (fuentes homogéneas de datos), modelos similares a los descubiertos al analizar los datos de ambos conjuntos centralizados en D. Las conclusiones al respecto a las que se arribaron fueron las siguientes:

- Con el algoritmo de clasificación J48 las reglas de clasificación obtenidas a partir de cada conjunto de entrenamiento (D, D1 y D2) tienen muy pocos elementos en común. Esto

se explica porque dicho algoritmo es muy sensible a los cambios en los datos del conjunto de entrenamiento. Por ello, en casos como este, por ejemplo, puede ser complicado determinar, a partir de la "integración" de modelos de minería de datos correspondientes a fuentes de datos homogéneas, un modelo de minería de datos válido para el conjunto formado por la unión de todas esas fuentes de datos. En estos casos, para tener éxito en el proceso de "integración", parece ser necesario auxiliarse de otras técnicas o disponer de algún tipo de información adicional sobre las diferentes fuentes.

- Con el algoritmo Apriori las reglas de asociación obtenidas a partir de los conjuntos de entrenamiento disjuntos D1 y D2, aunque con valores de soporte y confianza diferentes, son similares a las obtenidas en el conjunto de datos centralizados (D). Por ello, en casos como este, es posible determinar, a partir de la "integración" de modelos de minería de datos correspondientes a varias fuentes homogéneas, un modelo de minería de datos válido para el conjunto formado por la unión de dichas fuentes.

Conclusiones

En este trabajo se ha desarrollado un estudio experimental, mediante la aplicación de técnicas de minería de datos, para el análisis de las coronariografías realizadas a pacientes con cardiopatía isquémica. Como resultado, se confirmó la hipótesis inicial de que la minería de datos facilita las investigaciones científicas sobre el tema en la especialidad de cardiología. Se obtuvo un modelo de clasificación (reglas de decisión) de las coronariografías, identificando las características que afectan el comportamiento de la variable complicaciones en este tipo de proceder quirúrgico. Además, se identificaron asociaciones importantes entre los factores de riego presentes en los pacientes con esta patología. Se realizó un análisis comparativo entre los resultados del análisis de datos centralizados y del análisis de datos distribuidos homogéneos.

Referencias

[1] Hernández, J., Ramírez, M. and Ferri, C. Introducción a la Minería de Datos, Pretice Hall, Madrid, 2004.

[2] Fayyad U, Piatesky-Shapiro G, Smyth P, Uthurusamy R. Advance in Knowledge Discovery and Data Mining. MIT Press, Cambridge, Mass, 1996.

[3] Sangüesa R, Molina LC. Data Mining, una introducción. Ediciones UOC, First Edition, 2000.

[4] Weka Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka. (Consultada: Mayo 2008).

[5] Sierra B. Aprendizaje Automático: conceptos básicos y avanzados. Aspectos prácticos utilizando el software Weka, Prentice Hall, Madrid, 2006 (ISBN: 848322318X).

[6] Quinlan J. C4.5: Programs for machine learning, Morgan Kaufmann Pub., 1993 (ISBN: 1558602380).

[7] Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules in Large Databases. Proceedings of the 20th International Conference on Very Large Data Bases, pp 478-499, 1994 (ISBN 1558601538).